SNIA DEVELOPER CONFERENCE
SDC 25
By Developers FOR Developers

Hyatt Regency Santa Clara, CA
September 15-17, 2025

Cloud Storage considerations for RAG in AI applications

Scott Hoag
Principal Product Manager
Azure Storage

www.sniadeveloper.org

Agenda

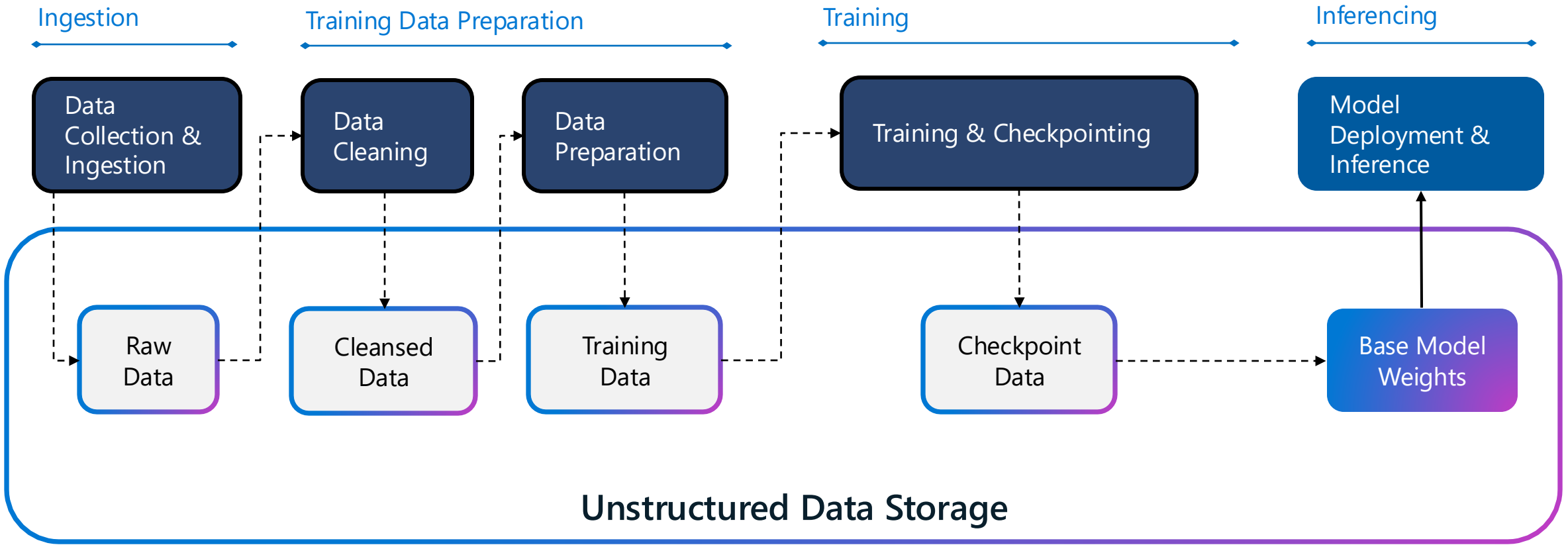
AI Workloads and Storage requirements

Retrieval Augmented Generation (RAG)

Microsoft

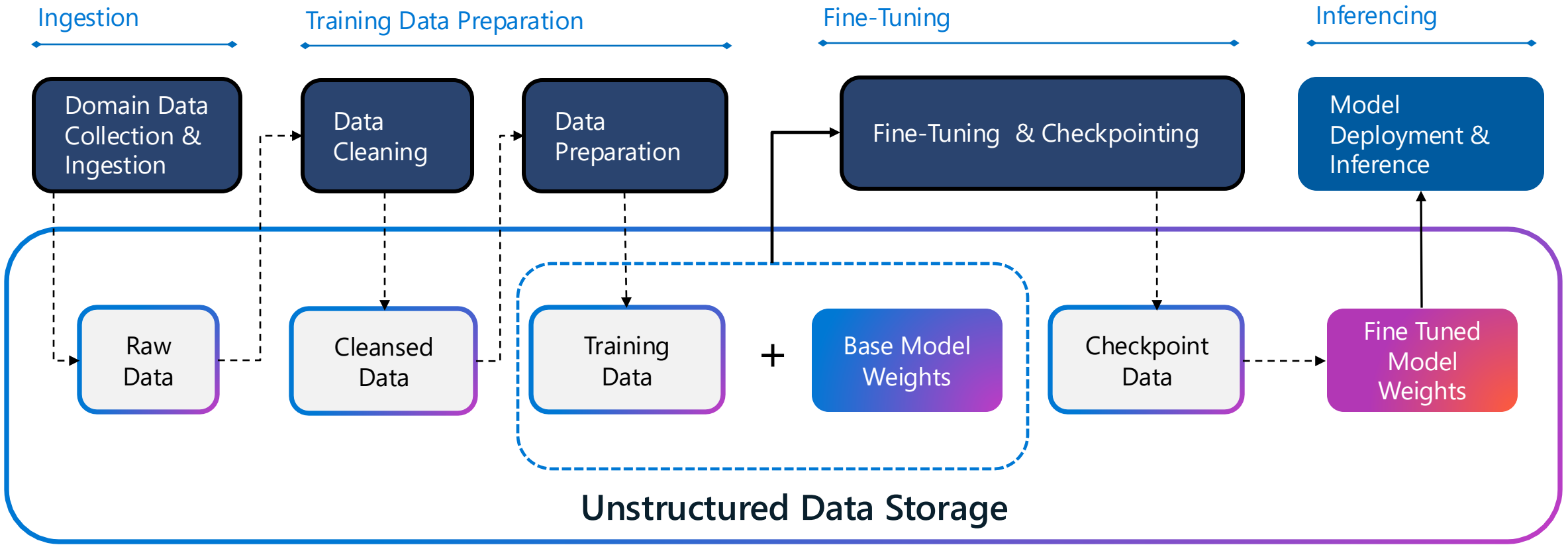
AI Workloads & Storage Requirements

AI Pipeline – Storage centric view



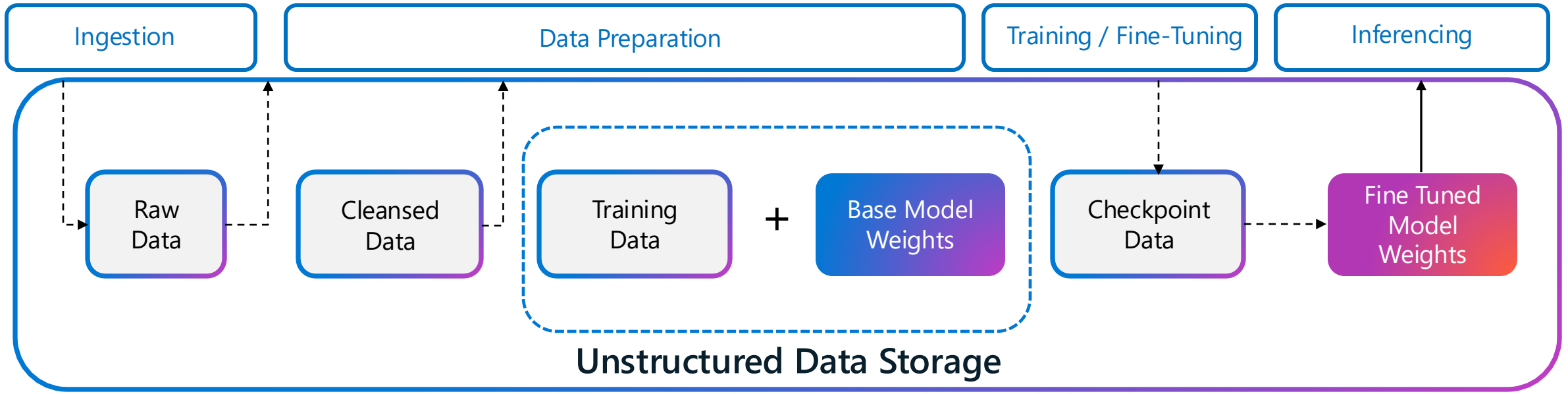
Microsoft

AI Fine-Tuning – Storage centric view



Microsoft

AI Pipeline - Storage Requirements



Requirements

Training / Fine-Tuning

- **Ingestion:** Bring raw training data to Azure
- **Data Preparation:** Integration with Spark, MosaicML, etc.
- **Training/Fine-Tuning:** Data to GPU nodes, checkpoints to storage. Integration with PyTorch and other ML frameworks
- **Data Management:** Secure & cost-efficient retention

Deployment/Inference

- **Deployment:** Model distribution and load times
- **Data Management:** Model versioning, retention of inference inputs and outputs

Microsoft

Azure Storage portfolio

Durable, highly available, massively scalable

Block storage

Services
- Azure Disk Storage
- Azure Elastic SAN
- Azure Container Storage

Unique capabilities
- Azure Elastic SAN
- Zonal Redundancy; Shared Disks

Object storage

Services
- Azure Blob
- Azure Data Lake Storage

Unique capabilities
- Premium Blob
- Multi protocol Access (e.g., NFS, HDFS)

File Storage

Services
- Azure Files
- Azure NetApp Files
- Azure Managed Lustre

Unique capabilities
- Native NetApp File Storage
- Azure File Sync

Data Management & Migration

New **Azure Storage Actions**

New **Storage Discovery** Preview

Storage Mover

Azure Data Box

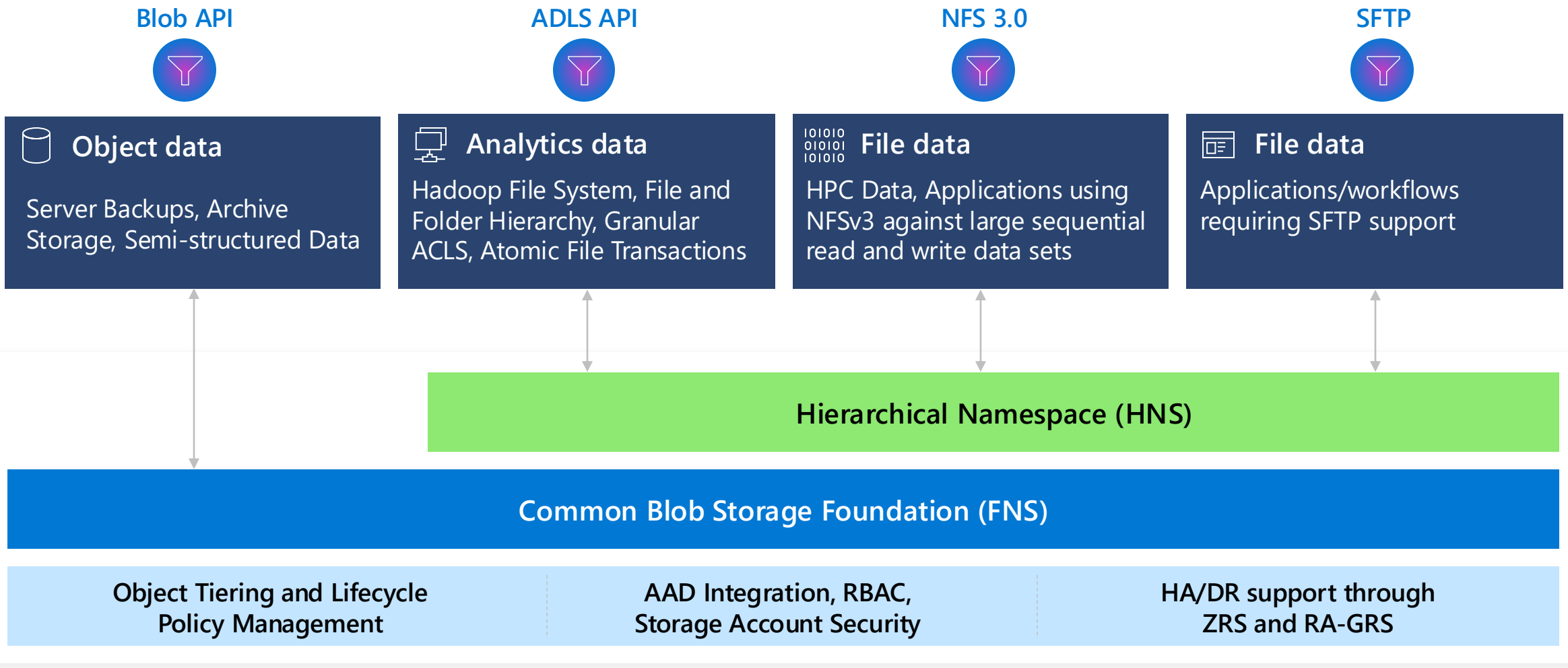
Capacity
100s of trillions of objects across many exabytes of data

Throughput
>600 Tbps average
(>200 exabytes per month)

IOPS
>700M tps
(~2 quadrillion per month)

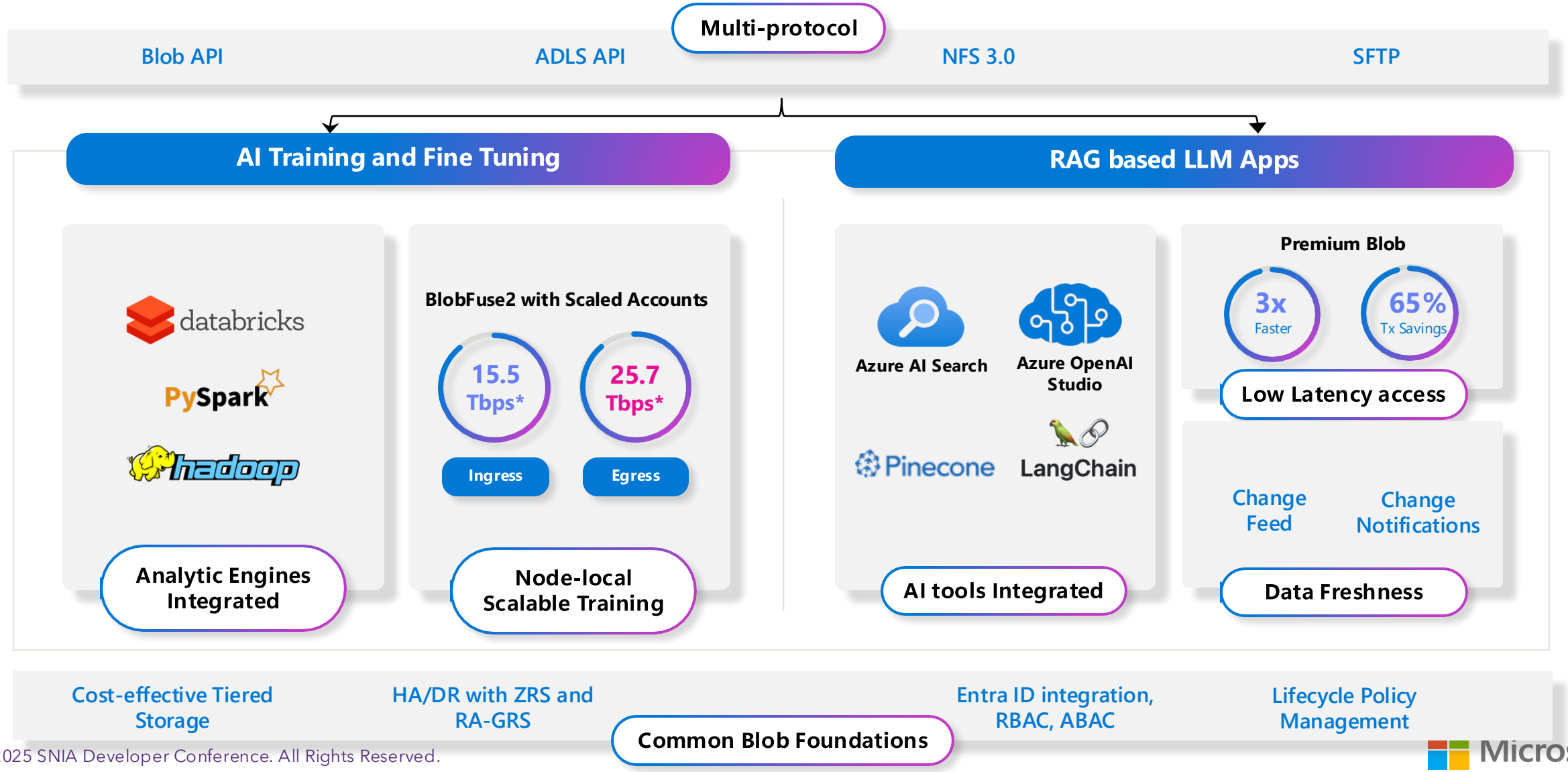
Microsoft

Azure Blob Storage: Multi-protocol, single storage platform



Microsoft

Build AI Apps with Azure Storage



Microsoft

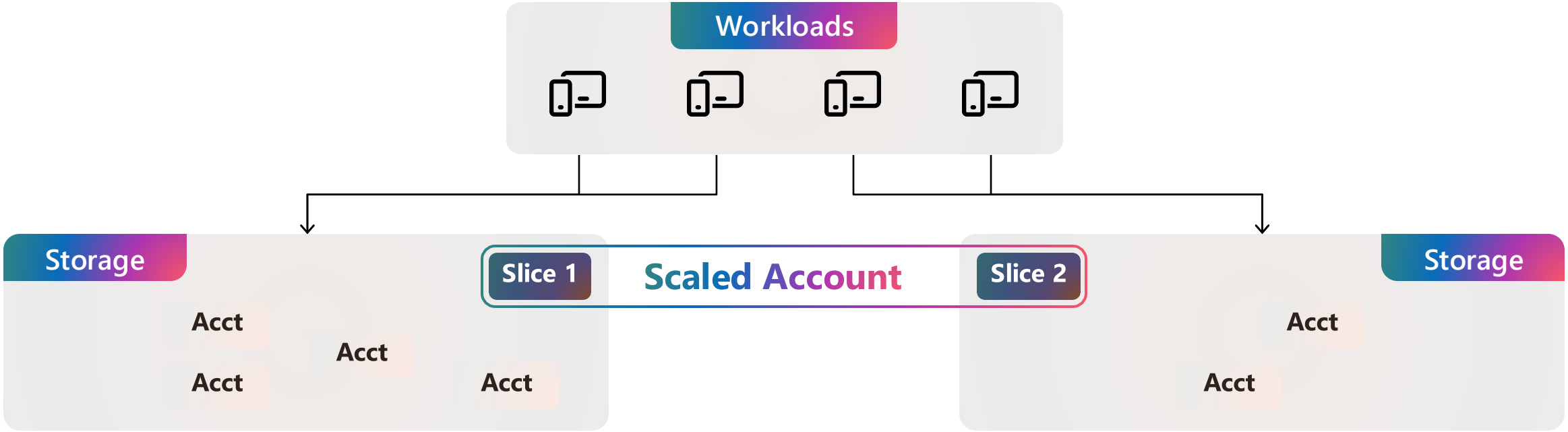
Blob Storage: Scaled accounts

Increase Capacity, Bandwidth, IOPS to meet hyperscale customer asks

Scale across a much larger hardware footprint

Eliminate the need for sharding large workloads across accounts

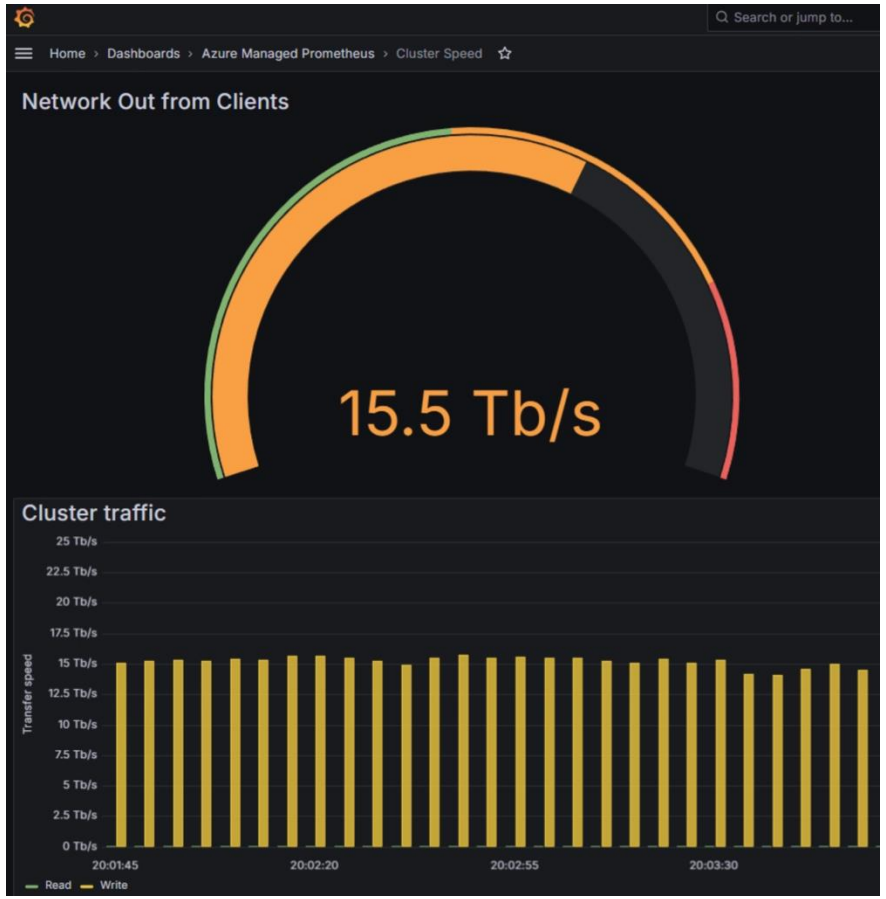
No change to pricing, not a new product/SKU, goal is to auto-scale



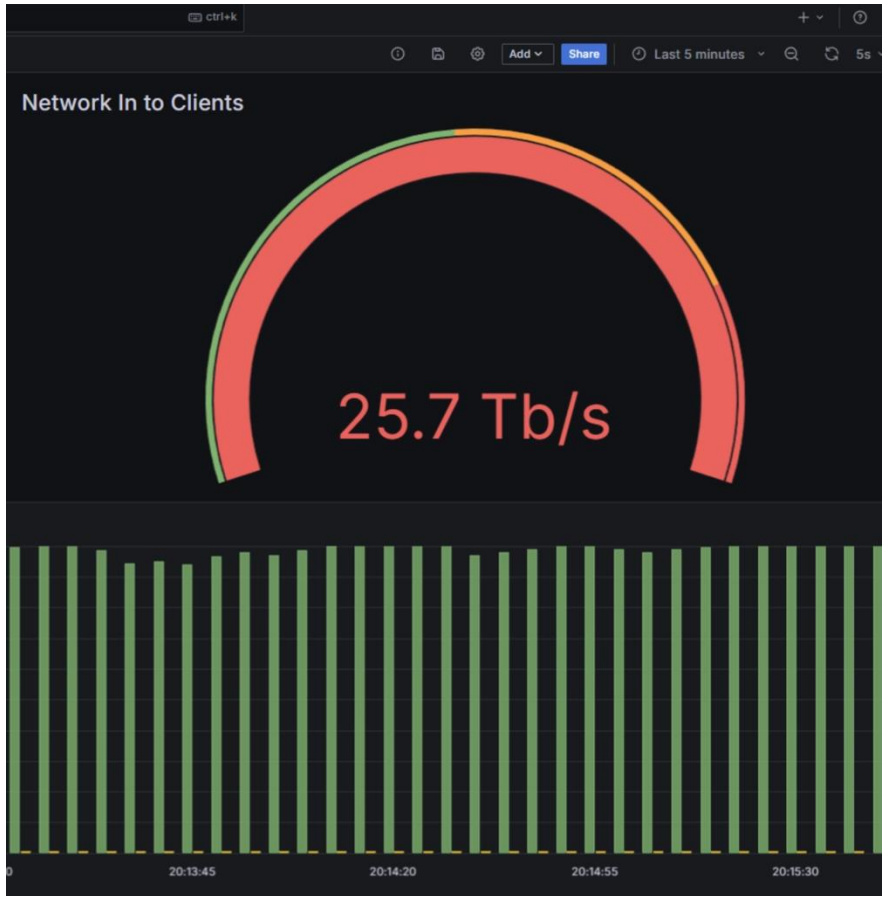
Microsoft

HPC IOR Benchmark test for Storage

Writing to Storage Account



Reading from Storage Account



Microsoft

Enabling data movement with AzCopy

AzCopy is a “Swiss Army Knife” command-line utility developed by Microsoft for transferring data to and from Azure Storage accounts

It is optimized for high performance, making it suitable for tasks such as uploading, downloading, copying, and synchronizing data across various storage services within the Azure ecosystem

Microsoft

BlobFuse2

High Throughput access to Blob Storage

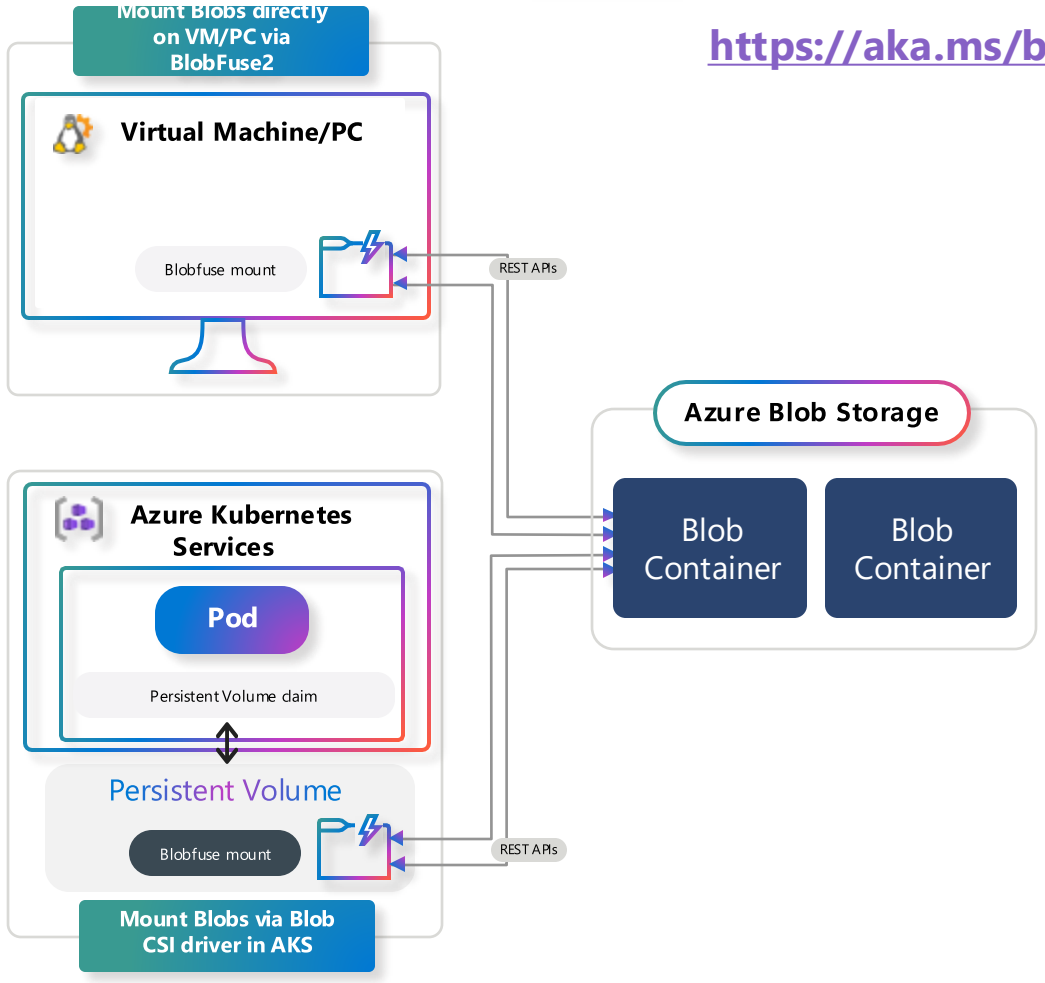
Easy to install and work with PiB scale data

Open-sourced & supported by Microsoft

Secure access to data

Virtual File System Driver to mount blob storage in your computing environment

https://aka.ms/blobfuse



Microsoft

Now available

Azure Storage Connector for PyTorch

Direct to Blobs

Python library offering PyTorch primitives to connect to Blob Storage

Integrated

Supports PyTorch datasets and loading & saving PyTorch models

Straightforward

Simple, directly pluggable integrations to PyTorch

Fast

Faster than other Azure Storage Python libraries

aka.ms/azstoragetorch

Microsoft

RAG with Object Storage

Bringing domain knowledge to LLMs

Prompt engineering

In-context learning

Fine-tuning

Learn new skills

Retrieval augmentation (RAG)

Learn new facts

Microsoft

RAG: Retrieval-Augmented Generation

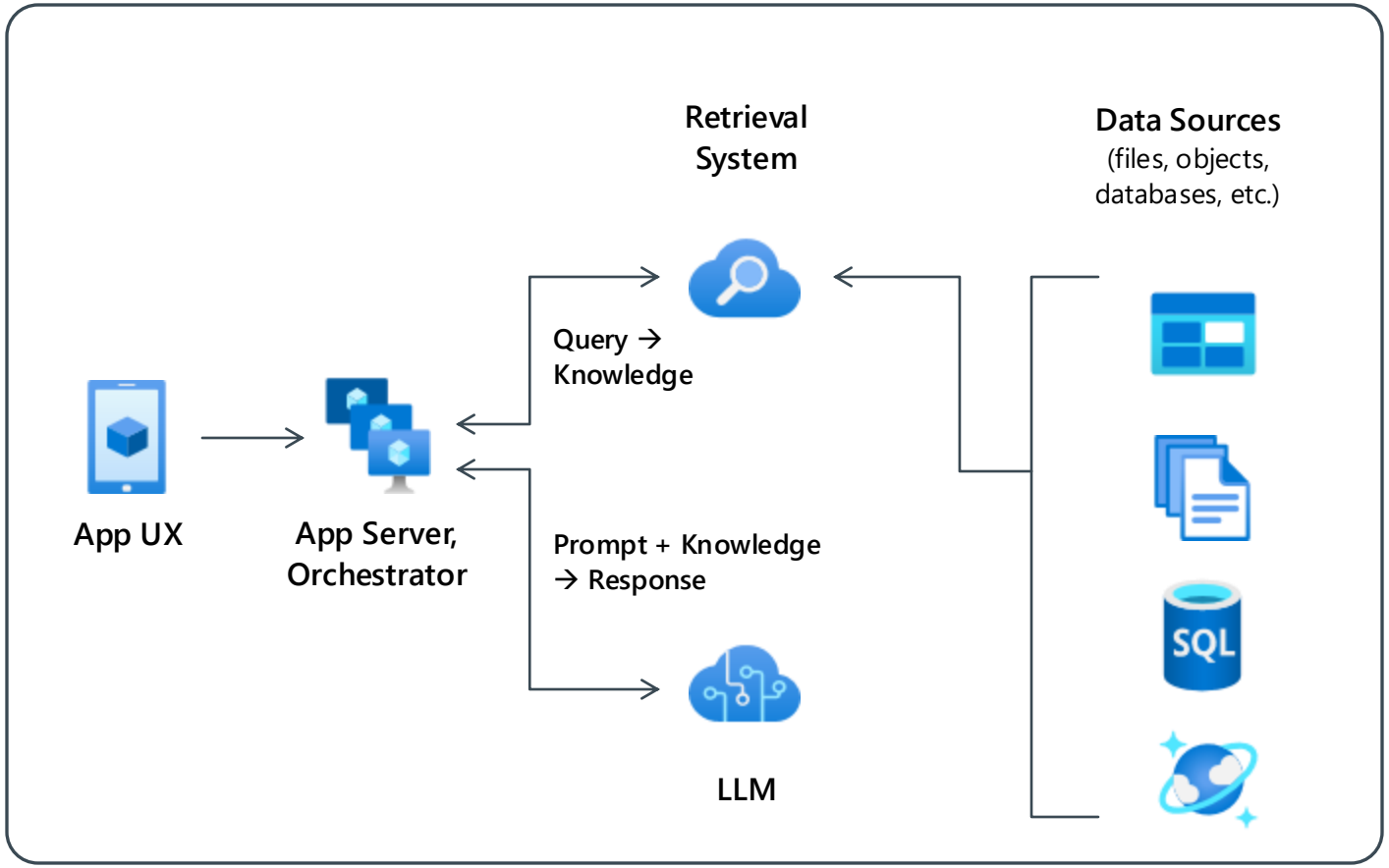
Combine reasoning + knowledge

Key Elements

- Externalized knowledge
- Orchestrator drives interaction
- Prompts =
  instructions + context + grounding data

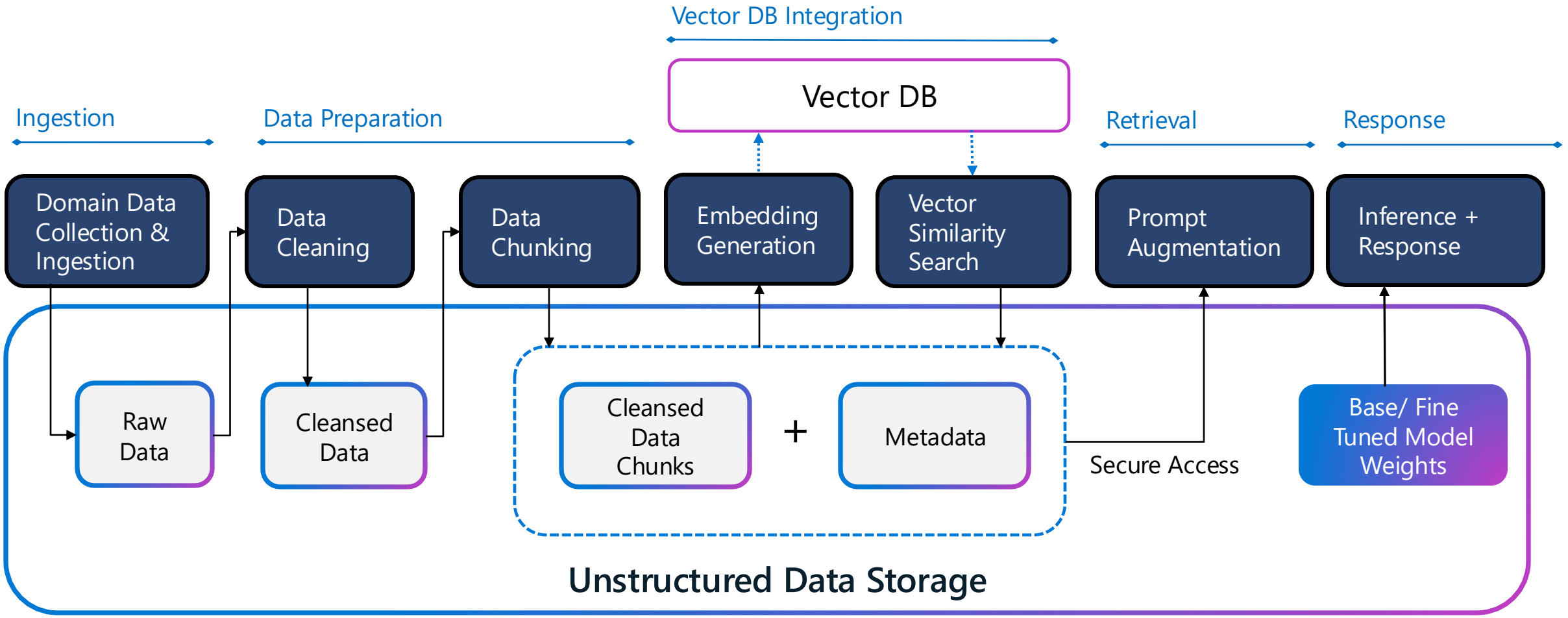
Achieving quality results

- Different workflows for different tasks
- Evaluation & RAI



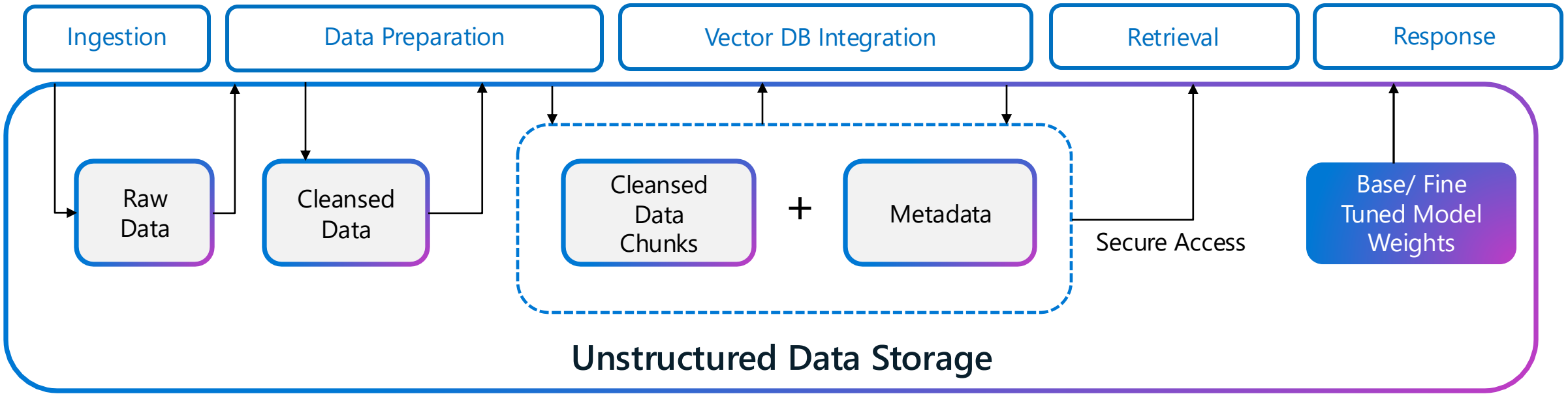
Microsoft

RAG Pipeline – Storage centric view



Microsoft

RAG Pipeline – Storage Requirements



Requirements

- Bring enterprise and domain-specific data to foundation models
- Low latency access to models and data
- Integration with Vector DB functionality
- Timely data and index updates
- Security and access control to core enterprise data

Microsoft

RAG with Blob Storage

Multi-Protocol

Unified storage for heterogenous updates

Low latency access

Premium Blob Storage

Vector DB Integration

Azure AI search, flexibility to BYO

Multiple indexes, dev-focused SDK/tools

Freshness

Blob Change Feed

Change notifications

Security

Azure Entra ID integration

RBAC and ABAC

Microsoft

Accelerating Performance: SSD-backed Object Storage

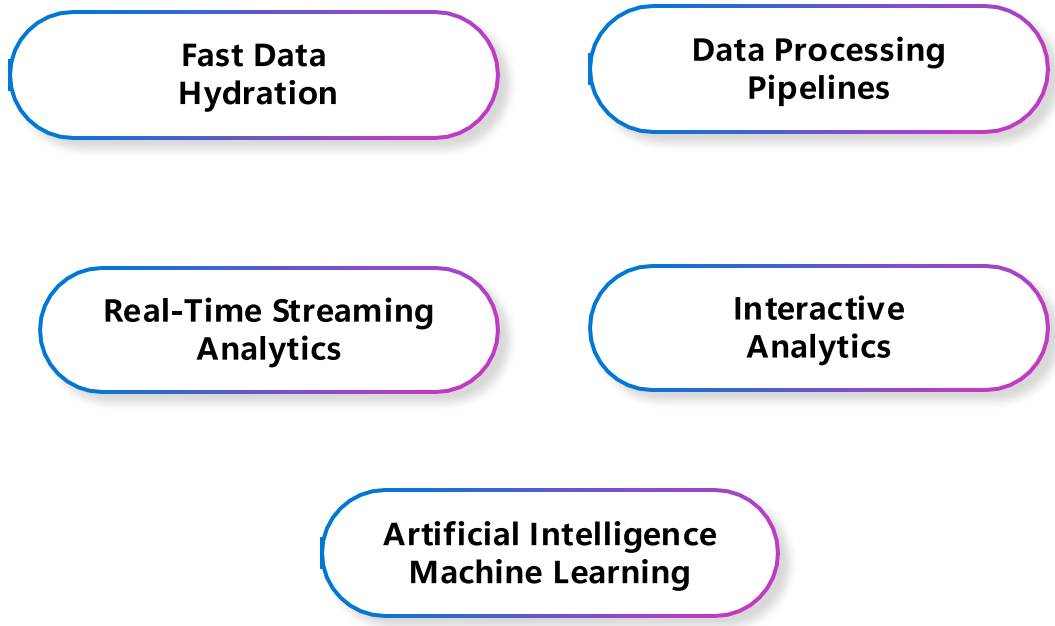
SSD-backed object storage unlocks unprecedented performance, revolutionizing data-intensive applications.

Low Latency
SSDs offer significantly lower latency compared to HDDs, enabling faster data retrieval for smaller objects/IOs.

High IOPS
SSDs handle much higher IOPS, enhancing performance for transactional workloads

High Throughput
SSDs provide higher data transfer rates, crucial for large-scale data processing.

Consistent Performance Under Load:
SSDs maintain steady performance levels even under heavy workloads.



Workloads enabled by SSD-backed Azure Premium Blob Storage

Microsoft

SSD-backed Premium Blob Storage for RAG

	Latency (400KiB PDFs)
Operation	**Premium Blob (SSD) vs Standard (Hot) Blob (HDD)**
PutBlob (Data Ingest + Chunk Writes)	**Premium ~2x faster**
GetBlob (Vectorization)	**Premium ~3x faster**
GetBlob (Retrieval)	**Premium ~3x faster**

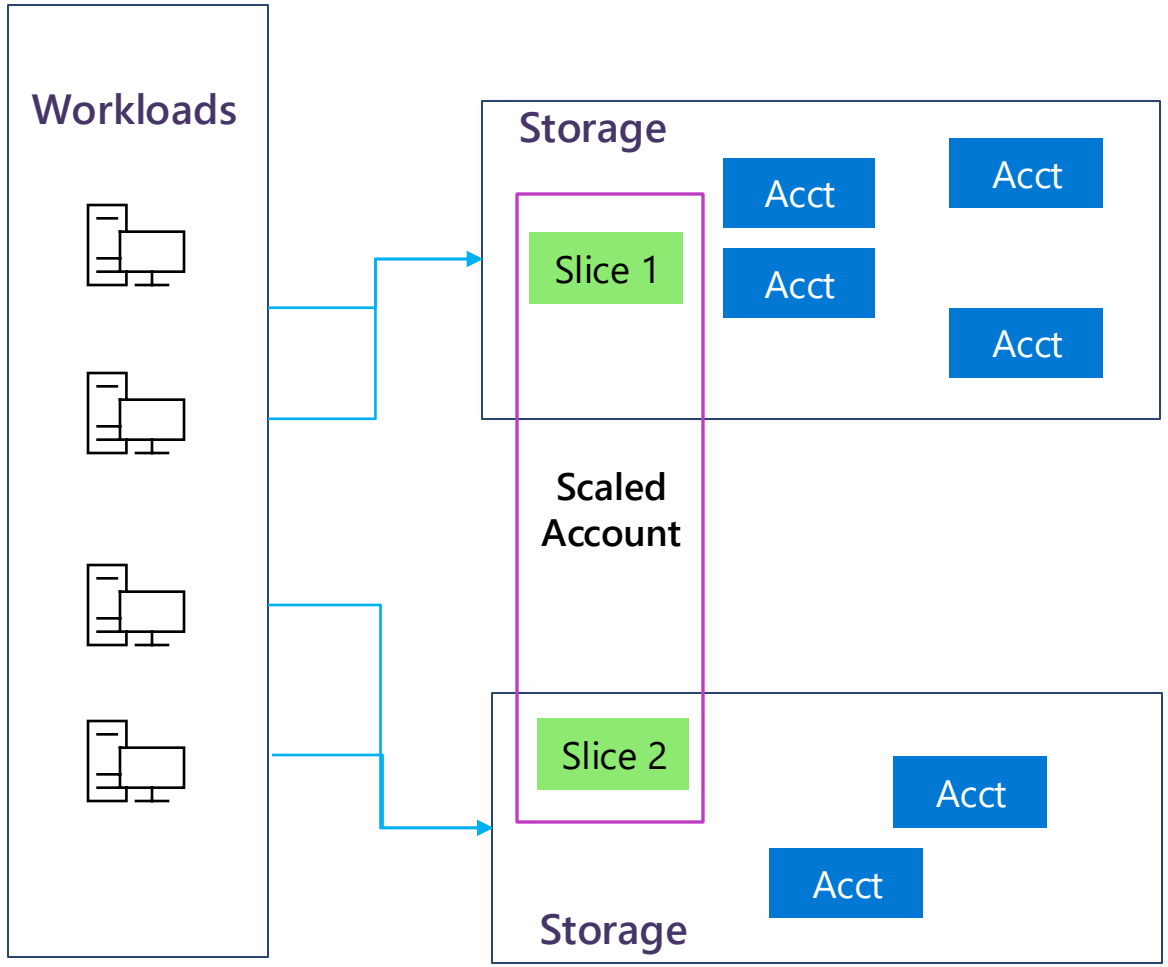
Premium delivers ~3x faster RAG performance with 65% savings on Transactions!

Microsoft

Present: Powering AI Pipelines and Vector DBs

Capabilities leveraged by AI Pipelines

- Disaggregated and scaled-out Storage Architectures (e.g. Blob Storage Scaled Accounts)
- GPU Node-local access and caching (BlobFuse2)
- Analytics engine integration via Multi-protocol access
- Lifecycle Management of Training/Inference Data



Azure Blob Storage Scaled Account

Microsoft

Present: Powering AI Pipelines and Vector DBs

Capabilities leveraged by AI Pipelines

- Disaggregated and scaled-out Storage Architectures (*e.g.,* Blob Storage Scaled Accounts)
- GPU Node-local access and caching (BlobFuse2)
- Analytics engine integration via Multi-protocol access
- Lifecycle Management of Training/Inference Data

Enabling Vector DB implementations

- Object Storage ideal to offload Vector DB indexes (*e.g.,* Pinecone) and chunks (*e.g.,* Azure AI search)
- Change Feed/Notifications ideal for logging dynamic changes to knowledge base (freshness)
- Enhanced Security with RBAC/ABAC

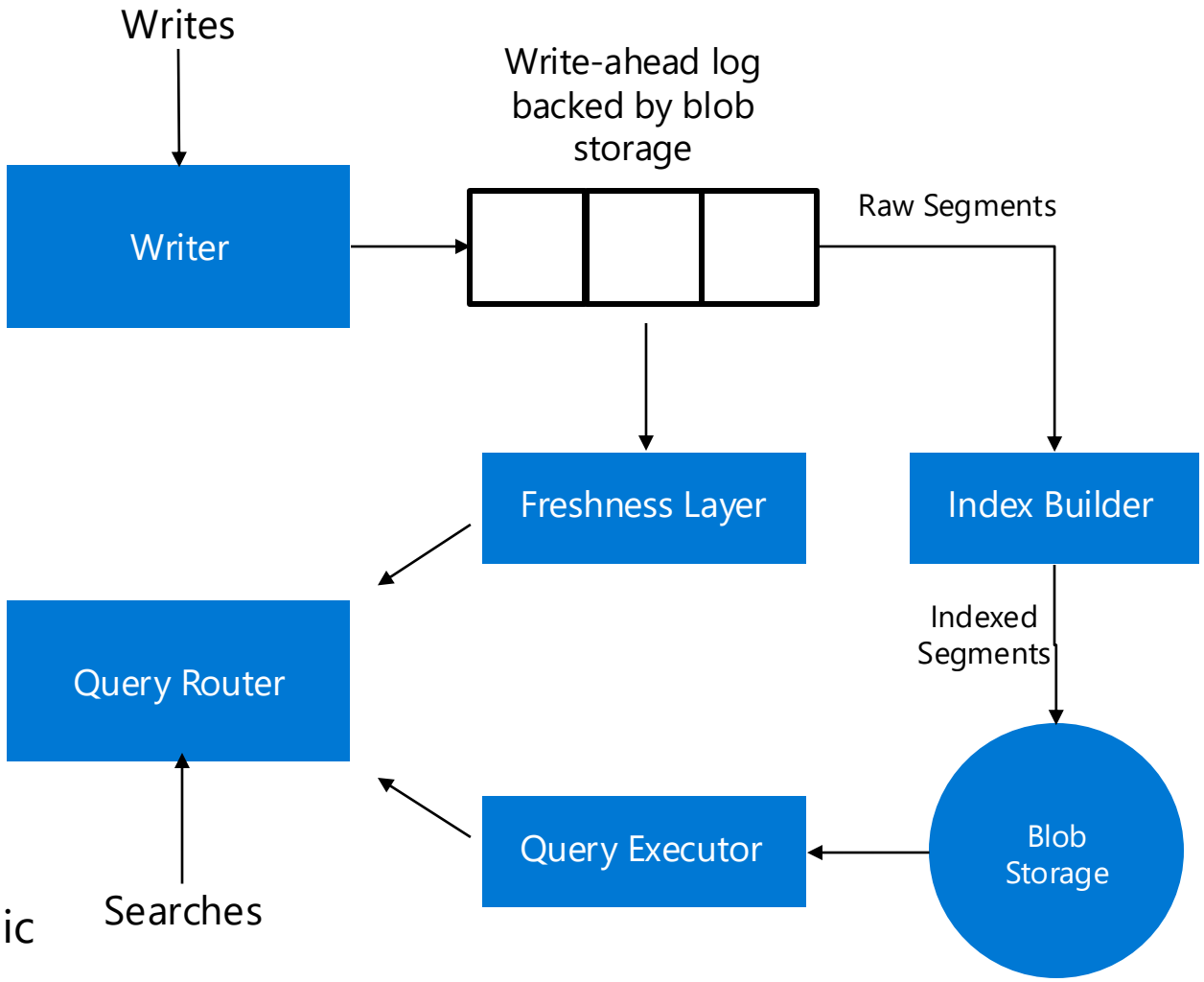


Image courtesy: Pinecone, source: [Reimagining the vector database to enable knowledgeable AI | Pinecone]

Microsoft

Key Takeaways

Leveraging Object Storage for building AI Apps …

Ideal for AI Training and Fine Tuning

Scalable

to Exabytes of data and many Tbps of throughput

Cost-effective

with storage tiers and automated lifecycle management

Integrated

with analytics engines for data preparation

Interoperable

GPU Node-local POSIX-like mount points

Accelerates building RAG based LLM Apps

Secure

with identity-based Authn/Authz,, RBAC

Interoperable

with vector DBs and orchestrators with SDK/tools

Low-Latency Access

with SSD-backed object storage

Freshness

with change feed and change notifications

Microsoft

Q&A

Please remember to rate this session. You get access the presentations at
https://sniadeveloper.org/conference

Microsoft